View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository of Brest State Technical University

Вестник Брестского государственного технического университета. 2009. №5
УДК 004.932.72'1; 004.93'14

Бушенко Д.А., Садыхов Р.Х.

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ ЛЕНТОЧНОЙ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ РАЗДЕЛЕНИЯ ПЕРЕСЕКАЮЩИХСЯ ПРОТЯЖЕННЫХ ОБЪЕКТОВ

Введение. Задача кластеризации часто возникает при разработке как прикладного, так и научного программного обеспечения. Но возможность применить без изменения классические методы кластеризации появляется довольно редко. Это связано с тем, что базовые алгоритмы кластеризации разработаны для идеальных исходных данных. В реальных же задачах приходится так или иначе модифицировать эти алгоритмы для адаптирования их к реальным данным. И хотя на сегодняшний день существует множество модификаций алгоритмов кластеризации, они не показывают достаточную эффективность при кластеризации пересекающихся протяженных объектов, что и стало причиной разработки нового модифицированного алгоритма кластеризации.

Задача кластеризации пересекающихся протяженных объектов появилась в процессе разработки цифровой экспертной системы идентификации текстильных волокон. Эта система выполняет следующие шаги в процессе идентификации:

1. Получение изображения со сканера;
2. Отделение волокон от фона;
3. Разделение пересекающихся волокон;
4. Идентификация искомых волокон.

На первом этапе цифровые изображения образцов текстильных волокон получаются при помощи широко распространенного сканера. Особенностью таких изображений является их слабая контрастность вместе с высокой зашумленностью.

На втором этапе волокна отделяются от фона при помощи сложных алгоритмов сегментации [1], [2]. В результате такой сегментации волокна действительно отделены от фона, но не отделены друг от друга. Таким образом, задача разделения пересекающихся волокон возникает уже на третьем этапе работы цифровой системы идентификации текстильных волокон. На рис. 1 показан типичный пример пересекающихся текстильных волокон, подлежащих разделению.

Рис. 1. Пересекающиеся волокна

Из-за особенностей технологического процесса получения образцов текстильных волокон, объекты на изображении зачастую очень похожи. Так, на приведенном примере (рис. 1) сложно разделить эти волокна даже вручную, не говоря уже о том, чтобы ожидать хороших результатов от автоматической кластеризации в ее классическом варианте. Поэтому конкретизируем задачу кластеризации таким образом, чтобы было возможно найти достаточное по своей эффективности решение.

Когда эксперт выполняет поиск текстильных волокон, идентичных указанному, при помощи описанной выше цифровой системы, результатом поиска должны быть протяженные объекты, визуально похожие на текстильные волокна.

Следовательно, применительно к примеру, показанному на рис. 1, требуемый алгоритм кластеризации не обязательно должен разделить изображение на несколько волокон. В случае, если будет установлена однородность объекта на рис. 1, он так и останется одним объектом. Но если кластеризация все же будет проведена, то объект на рис. 1 должен быть разделен на кластера таким образом, чтобы визуально они были похожи на текстильные волокна.

Алгоритмы кластеризации. Классические методы кластеризации основываются на предположении о некоторой форме исследуемых классов объектов. Наиболее значимые типы классов следующие [3]:

1. Класс типа ядра. Все расстояния между объектами внутри класса меньше расстояния до любого другого объекта.
2. Класс со сгущением в среднем. Среднее расстояние внутри класса меньше среднего расстояния объектов класса до всех остальных объектов.
3. Класс типа ленты. Внутри класса для каждого объекта всегда найдется другой объект, расстояние до которого меньше, чем расстояние до любого из объектов других кластеров.

В [3] предлагается классификация алгоритмов кластеризации на эвристические и иерархические. Целью иерархических методов кластеризации является построение дендрограммы, объединяющей объекты на разных уровнях в различные кластера. Для наших целей больше подходят эвристические методы, способные разделить пространство объектов на однородные группы.

Для кластеризации объектов типа 1–2 базовым является алгоритм К-средних [4]. Существуют и другие алгоритмы, работающие на сходных принципах, например «Форэль» и метод потнециальных ям [3], но они показывают похожие результаты. В случае объектов типа 3 базовым является алгоритм связанных компонентов [3]. Рассмотрим их подробнее применительно к кластеризации пересекающихся текстильных волокон.

Алгоритм К-средних. Алгоритм К-средних состоит в следующем. Случайным образом выбираются К-объектов, которые считают центрами кластеров. Все остальные объекты присоединяются к ближайшему из этих центров. Затем центры кластеров пересчитываются как центры масс объектов их составляющих. Процедура повторяется до тех пор, пока кластера не стабилизируются.

Обычно алгоритм К-средних применяется для разделения кластеров по координатам x и y, если эти кластера имеют ярко выраженную эллиптическую форму, т.е. принадлежат к классам 1 или 2. Однако в случае протяженных объектов алгоритм К-средних не применим к координатам, т.к. текстильные волокна имеют протяженную форму. Этот алгоритм можно применить для исследования таких характеристик волокон, как значение тона, насыщенности или яркости. На рис. 2 показана гистограмма распределения яркостей пересекающихся волокон, приведенных выше на рис.1.

Как видно из гистограммы, нет явно заметных пиков кластеров таких, чтобы можно было сказать, что два волокна разделимы по этим пикам. Поэтому алгоритм К-средних разобьет все значения яркостей на два кластера приблизительно посередине. Результат такой кластеризации показан на рис. 3. Действительно, исследуемый объект разделен на два кластера, но совершенно неподходящим образом.

Еще один серьезный недостаток алгоритма К-средних в том, что изначально неизвестно количество кластеров, на которые нужно разбить исследуемый объект.

Бушенко Д.А., *аспирант Белорусского государственного университета информатики и радиоэлектроники.*
Садыхов Рауф Хосровович*, д.т.н., профессор, зав. кафедры электронно-вычислительных машин Белорусского государственного университета информатики и радиоэлектроники.*
Беларусь, БГУиР, 224017, 220013, г. Минск, ул. П. Бровки, 6.

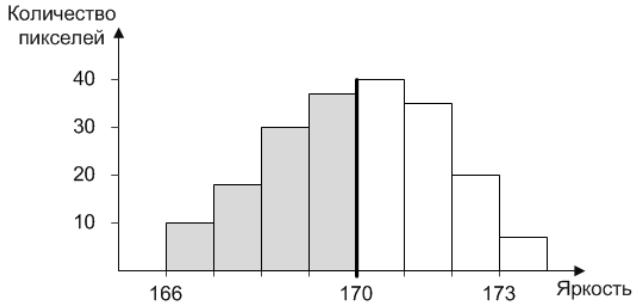

Рис. 2. Распределение яркостей

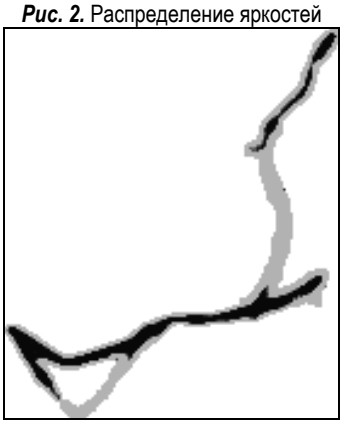

Рис. 3. Кластеризация яркостей методом К-средних

Алгоритм ленточной кластеризации. Алгоритм ленточной кластеризации, или алгоритм связных компонентов [3], состоит в следующем. Все объекты разбиваются на классы типа «лента» таким образом, чтобы для каждого объекта кластера всегда нашелся еще один объект из этого же кластера, расстояние до которого меньше, чем расстояние для любого из объектов других кластеров. При этом минимальное расстояние между объектами разных кластеров задается как t из $(\min d_i, \max d_i)$, где $\min d_i$ и $\max d_i$ – минимальное и максимальное расстояние между объектами соответственно. Блок-схема алгоритма приведена на рис. 4.

Алгоритм ленточной кластеризации предоставляет большие возможности по разделению изображений пересекающихся волокон в силу того, что он изначально предназначался именно для протяженных объектов. Следующие характеристики пикселей можно использовать для вычисления функции расстояния как по отдельности, так и совместно:

- Значение тона;
- Насыщенность;
- Яркость;
- Координаты x и y.

Это неоднородные характеристики, которые имеют совершенно различные свойства. Рассмотрим их подробнее.

Значение тона вычисляется в полярных координатах. В цифровой системе он зависит от «разрешающей способности» цветового круга [5]. Так, в описываемой цифровой системе идентификации текстильных волокон используется 360 значений тона по числу градусов в круге.

Насыщенность – это длина радиуса круга, в котором вычисляется значение тона. Значение насыщенности в описываемой цифровой системе имеет границы [0; 255].

Рис. 4. Алгоритм ленточной кластеризации

Расстояние между двумя точками в координатах тон/насыщенность может быть вычислено по формуле [6]:

$$d = \sqrt{s_1^2 + s_2^2 - 2s_1 s_2 \cos(h_2 - h_1)}, \qquad (1)$$

где S_1, S_2 – насыщенности двух точек, h_1, h_2 – значения тона двух точек.

Значение яркости также изменяется в пределах от 0 до 255.

В функции расстояния между пикселями будут использоваться координаты точки x и y, а также еще одна из описанных выше характеристик. Для того, чтобы определиться, какую из характеристик (тон, насыщенность или яркость) выбрать, посмотрим, как они изменяются на различных изображениях (таблица 1).

Таблица 1

Сканер	Изображение	Количество изображений	Характеристика	СКО[1], %
Epson Perfection 4870	**Глянцевая подложка, изображения размерами 5x10 см.**	**20**	Тон **Насыщенность** **Яркость**	15,64 **2,07** **1,16**
Epson Perfection 4870	**Подложка из чистой белой бумаги, изображения размерами 5x10 см.**	**20**	Тон **Насыщенность** **Яркость**	21,53 **1,77** **2,21**
Imacon	**Глянцевая подложка, изображения размерами 5x10 см.**	**20**	Тон **Насыщенность** **Яркость**	14,26 **2,31** **3,87**
Epson Perfection 4870	**Изображения волокон площадью более 100 пикселей без фона**	**2688**	Тон **Насыщенность** **Яркость**	4,43 **3,61** **4,33**

1) СКО – среднеквадратичное отклонение

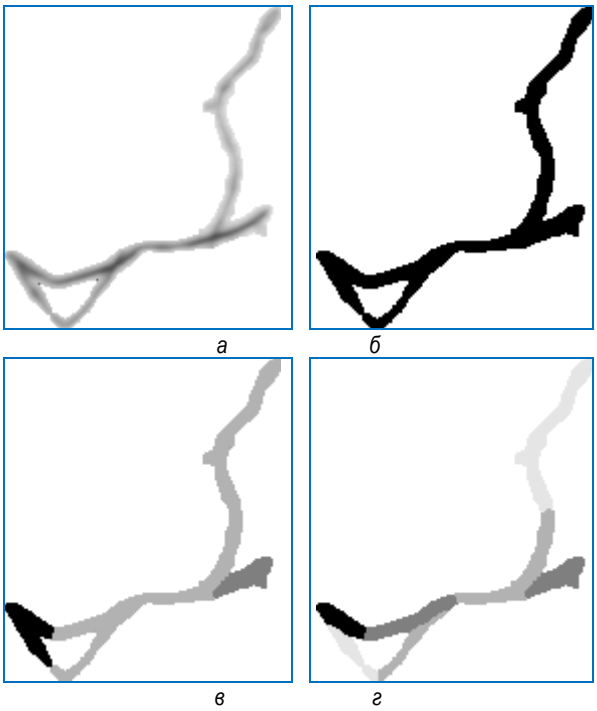

Рис. 5. *а* – исходное изображение (канал тона); *б* – кластеризация при *w* = 1, t = 10; *в* – кластеризация при *w* = 2, *t* = 10; *г* – кластеризация при *w* = 2, *t* = 20

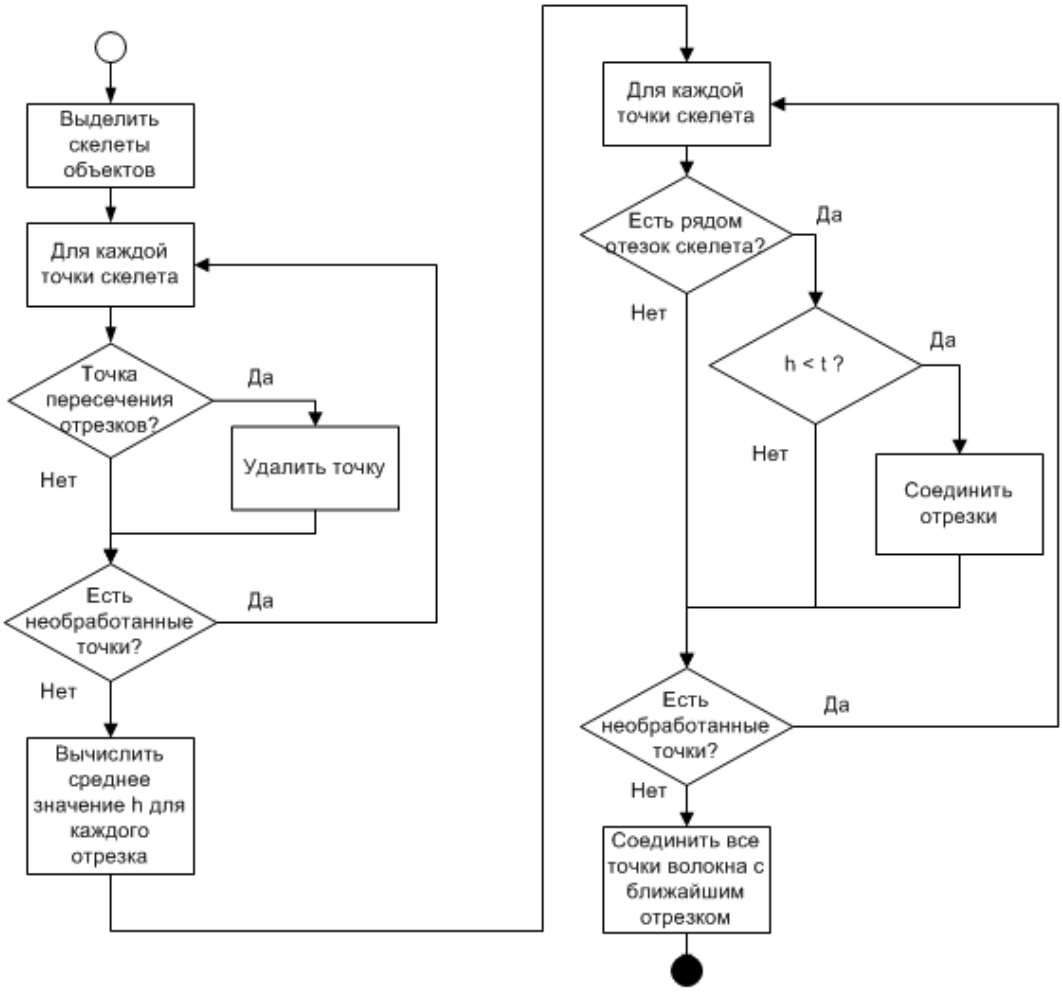

Рис. 6. Модифицированный алгоритм ленточной кластеризации

Как видно из результатов экспериментов, приведенных в таблице выше, СКО характеристик сильно зависит от условий, в которых получались изображения. Однако наибольшим СКО обладает значение тона и, следовательно, является наиболее информативным признаком. Поэтому в формуле расстояния между пикселями вместе с координатами x и y будем использовать дополнительно h – значение тона.

Поскольку координаты точки в пространстве [x, y, h] не обладают однородными свойствами, невозможно применить евклидово расстояние для вычисления расстояния между объектами. Воспользуемся в этих целях расстоянием Минковского [7]:

$$d_{ij} = d\left(X_i, X_j\right) = \left(\sum_{k=1}^{m} w_k \left|x_{ik} - x_{jk}\right|^q\right)^{1/q}, \quad (2)$$

где w_k>0 – весовые коэффициенты при признаках; $q \geq 1$ – параметр степени. Веса выражают степень важности измерения при различении объектов. При сопоставлении объектов измерениям признака с меньшей значимостью нужно присваивать меньшие веса.

Таким образом, применяя расстояние Минковского к пространству признаков [x, y, h] и приняв коэффициент значимости при координатах x и y за 1, а $q = 2$, получаем формулу:

$$d_{ij} = d\left(X_i, X_j\right) = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + w\left(\overline{h_i - h_j}\right)^2}, \quad (3)$$

$$\overline{h_i - h_j} = \overline{d\left(H_i, H_j\right)} =$$

$$= \begin{cases} \left|h_i - h_j\right|, \text{ при } \left|h_i - h_j\right| < h_{\max}/2 \\ h_{\max} - \left|h_i - h_j\right|, \text{ в остальных случаях}, \end{cases} \quad (4)$$

где $h_{\max}$ – максимально допустимое значение тона.

Из (3) видно, что для правильного определения расстояния между объектами в пространстве [x, y, h] ключевым является верный выбор коэффициента w.

На рис. 5 показаны результаты ленточной кластеризации пересекающихся протяженных объектов с применением функции расстояния (3).

Из рис. 5 видно, что результаты кластеризации сильно зависят от обоих заданных параметроов – w и t. Однако даже в таком виде качество кластеризации значительно выше в сравнении с кластеризацией методом К-средних (рис.3). Выделенные кластера представляют собой объекты протяженной формы, визуально похожие на текстильные волокна.

Показанный выше алгоритм ленточной кластеризации с применением функции расстояния (3) для пространства координат [x, y, h] имеет следующие недостатки:

1. Сложность правильного выбора параметров w и t. Эмпирический выбор здесь не подходит в виду того, что значение параметра h сильно изменяется в разных условиях (см. табл. 1). В случае неверного выбора параметров w и t, результат кластеризации будет неудовлетворительным.
2. В случае, если гистограмма распределения яркостей, показанная на рис.2, будет иметь достаточно крутые склоны, и расстояние пикселей по краям волокна до пикселей в центре волокна будет больше t, то возникнет эффект «вложенных» объектов, как на рис. 3.

Модифицированный алгоритм ленточной кластеризации. Для преодоления указанных недостатков был разработан модифицированный алгоритм ленточной кластеризации. Блок-схема алгоритма приведена на рис. 6.

На первом шаге работы алгоритма извлекаются скелеты протяженных объектов. Существуют различные сложные методы извлечения скелетов, например такие, как представлены в [8], [9]. Однако наиболее простым для реализации и достаточно эффективным для наших задач является алгоритм утончения, показанный в [10] и [11].

На втором шаге алгоритма все скелеты делятся на отрезки таким образом, чтобы в них не было пересечений. Условие, которому должен соответствовать каждый пиксель такого отрезка: у этого пикселя должно быть не более двух соседних пикселей скелета.

Затем вычисляется среднее значение тона каждого отрезка скелета как медиана распределения тона [3].

Следующий шаг – алгоритм машинной графики «заполнение контура по критерию связности» [12], соединяя соседние отрезки, у которых разность значений тона, вычисленная по формуле (4), не превышает заданного значения t. Каждый соединенный набор отрезков относится к уникальному кластеру.

На последнем шаге все точки объекта заносятся в тот кластер, которому принадлежит их ближайший отрезок. Результат кластеризации показан на рис. 7.

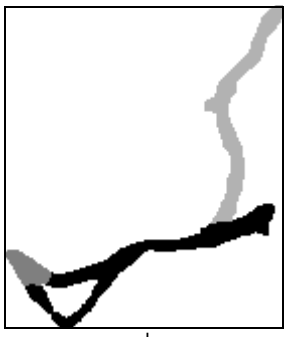

Рис. 7. Обработка методом модифицированного алгоритма ленточной кластеризации (t = 10)

Как видно из рис. 7, результат кластеризации – объекты, визуально похожие на текстильные волокна, что полностью соответствует поставленной задаче в начале статьи.

Отличия модифицированного алгоритма от его классического аналога носят качественный характер. Они сведены в таблицу 2.

Таблица 2

Характеристика алгоритма	**Классический алгоритм ленточной кластеризации**	**Модифицированный алгоритм ленточной кластеризации**
Количество параметров, регулирующих работу алгоритма	2	1
Возможность получения объектов, непохожих на текстильные волокна	да	нет
Объект, которым манипулирует алгоритм	пиксель	отрезок скелета
Функция расстояния	Минковского	Евклидово

Заключение. Существуют множество алгоритмов кластеризации, предназначенных для различных типов объектов. Однако ни один из них не показывает удовлетворительных результатов при разделении пересекающихся протяженных объектов, таких как текстильные волокна. Алгоритм ленточной кластеризации, наиболее оптимизированный для протяженных объектов, сложен в реализации из-за того, что регулируется двумя параметрами, а его качество сильно зависит от формы гистограммы распределения тона объекта. Поэтому была разработана модификация алгоритма ленточной кластеризации, адаптированная специально для разделения пересекающихся протяженных объектов. Она регулируется всего лишь одним параметром, что значительно упрощает реализацию алгоритма, и инвариантна относительно формы гистограммы распределения тона объекта. В результате кластеризации предложенным алгоритмом получаются протяженные объекты, визуально похожие на текстильные волокна. Разработанный алгоритм является наиболее подходящим для кластеризации пересекающихся протяженных объектов, таких, как текстильные волокна.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Бушенко, Д.А. Модифицированный алгоритм адаптивной пороговой сегментации в задачах выделения протяженных объектов на слабоконтрастных изображениях / Д.А. Бушенко, Р.Х. Садыхов // Материалы IV международной конференции IST'2008. –, Минск, 2008. – С. 106–111.
2. Bushenko D.A., Sadykhov R.Kh. Segmentation of Extensive Objects on Low-Contrast Images // Proceedings of ICNNAI'2008, Minsk, 2008.

3. Лагутин, М.Б. Наглядная математическая статистика. – Москва, 2007.
4. Baldock R., Graham J., Image Processing and Analysis A Practical Approach, Oxford, 2000.
5. Levkowitz H., Color Theory and Modelling for Computer Graphics, Vizualization, and Multimedia Applications, Norwell, 1997.
6. Корн, Г. Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. – Москва, 1974.
7. Большаков, А.А. Методы обработки многомерных данных и временных рядов / А.А. Большаков, Р.Н. Каримов. – Смоленск, 2007.
8. Song Chun Zhu, Yuille A.L., FORMS: A Flexible Object Recognition and Modelling System, Harvard Robotics Lab. Technical Report no 94-101.
9. Gold C., Crust and Anti-Crust: A One-Step Boundary and Skeleton Extraction Algorithm, Quebec City.
10. Gonzales R.C., Woods R.E., Digital Image Processing, New Jersey, 2002.
11. Lam, L., Seong-Whan Lee, Ching Y. Suen, Thinning Methodologies-A Comprehensive Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No. 9, September 1992, page 879.
12. Павлидис, Т. Алгоритмы машинной графики и обработки изображений.

Материал поступил в редакцию 11.11.09

BUSHENKO D.A., SADYKHOV R.Kh. Modified algorithm of belt clustering for separation of crossing extended objects

There are various clustering algorithms which use different types of input data. The goal of the article is to develop a special clustering algorithm optimized for separation of crossing extended objects such as textile fibers. In this paper two types of algorithms are discussed and two most popular instances of these algorithms are observed: the C-means algorithm and the belt clustering algorithm. Since it is impossible to apply these algorithms for the task of separation of the crossed textile fibers, a special modification of the belt clustering algorithm is proposed. It is also presented a special space of descriptors whose effectiveness in separation of crossed extended objects is proved using the experimental results. Because of the fact that these descriptors contain nonuniform elements, it is also needed a special distance function. In this article it is proposed to use the Minkovsky distance. In conclusion, the comparison of the pure belt clustering algorithm and the modified one are discussed, and the advantages of the developed algorithm are shown in the task of separation of crossed extended objects.

УДК 004.8.032.26

Войцехович Л.Ю., Головко В.А., Курош Мадани

МУЛЬТИАГЕНТНАЯ СИСТЕМА ОБНАРУЖЕНИЯ АТАК С НЕЙРОСЕТЕВЫМ КЛАССИФИКАТОРОМ

Введение. Оперативный обмен информацией становится неотъемлемым атрибутом успешной деятельности в любой сфере. В последнее время прорыв в этой области обеспечили компьютерные технологии: компьютерные сети, электронная коммерция, корпоративные web-сайты и др. Однако наряду с необходимостью повышения надежности и скорости коммуникации остро встал вопрос обеспечения защиты информационных ресурсов.

Для защиты компьютерных систем применяются различные подходы. Все подходы можно разбить на две основные категории: организационные и технические. В свою очередь технические подходы подразделяются на сетевые и хостовые. Далее речь пойдет о сетевых средствах обеспечения безопасности, а именно, о системах обнаружения вторжений.

Задачей *Систем Обнаружения Вторжений* (*Intrusion Detection Systems – IDS*) является защита компьютерных сетей.

Наряду с правильной политикой безопасности, архитектурой межсетевых фильтров, антивирусным программным обеспечением и другими средствами IDS часто отводится роль основного элемента защиты. IDS используются в качестве средства раннего оповещения о сетевых проблемах. Это обусловлено размещением IDS в общей схеме обороны на сетевом уровне, на котором подозрительные действия могут быть обнаружены раньше, чем на более высоких уровнях. Кроме того, IDS способна предоставлять необходимые доказательства злоумышленных действий, а также выявлять скрытые тенденции, что становится возможным при анализе большого количества данных, обрабатываемых IDS.

К недостаткам существующих моделей IDS, в первую очередь, можно отнести уязвимость к новым атакам, низкую точность и скорость работы. Современные системы обнаружения вторжений плохо приспособлены к работе в реальном режиме времени, в то время как возможность обрабатывать большой объем данных в реальном времени – это определяющий фактор практического использования систем IDS. Указанные недостатки трудно устранить, используя только классические методы в области компьютерной безопасности. Поэтому в последнее время системы IDS активно изучаются. Разработчики систем обнаружения вторжений предлагают различные подходы: статистические методы [1, 2], нейронные сети [3, 4], деревья решений и SVM [5], генетические алгоритмы и искусственные иммунные системы [6, 7, 8, 9, 10].

В области обнаружения вторжений существует два основных метода: *обнаружение злоупотреблений* и *обнаружение аномалий*. Обнаружение злоупотреблений предполагает наличие сигнатур атак. Основным недостатком таких систем является их неспособность обнаруживать новые или неизвестные атаки, т.е. записи о которых в системе отсутствуют. Обнаружение аномалий [11] связано с построением профиля нормального поведения системы. При этом атакой считается любое отклонение от этого профиля. Главным преимуществом таких систем является принципиальная возможность определения ранее не встречавшихся атак.

Результаты исследований биологических механизмов *Иммунной системы человека* могут быть положены в основу построения систем обнаружения атак, поскольку базовые принципы работы в этих двух случаях схожи [12]. В иммунной системе человека имеются отдельные механизмы индивидуальной защиты и врожденного иммунитета, выполняющие функции аналогичные обнаружению злоупотреблений. Иммунная система человека состоит из различных иммунных клеток, химических сигналов, волокон и т.п. Их совместная работа позволяет обнаруживать определенные отклонения в организме человека, различать их и запускать необходимые механизмы иммунного ответа. А такие характеристики иммунной системы человека, как распределенность и самоорганизация (приспосабливаемость к изменчивым условиям), отвечают основным требованиям систем обнаружения аномалий. Таким образом, моделирование искусственной иммунной системы связано с разработкой алгоритмов динамического создания и обновления сигнатур, а так же алгоритмов обнаружения аномалий посредством сравнения с текущим состоя-

***Войцехович Леонид Юрьевич**, аспирант 3-го года обучения кафедры интеллектуальных информационных технологий Брестского государственного технического университета.*

***Головко Владимир Адамович**, д.т.н., профессор, зав. кафедрой интеллектуальных информационных технологий Брестского государственного технического университета.*

Беларусь, БрГТУ, 224017, г. Брест, ул. Московская, 267.

***Курош Мадани**, доктор наук, профессор университета Париж-XII.*

Франция, Париж, 12 Val de Marne (UPVM).